# Optimising performance of genAI in SLR screening using PICOS criteria

MSR201

Additional interactive content! Scan the QR code below

Automating screening using generative AI (GenAI) **could speed up literature reviews**, but performance may depend on model type, topic complexity, and prompt design

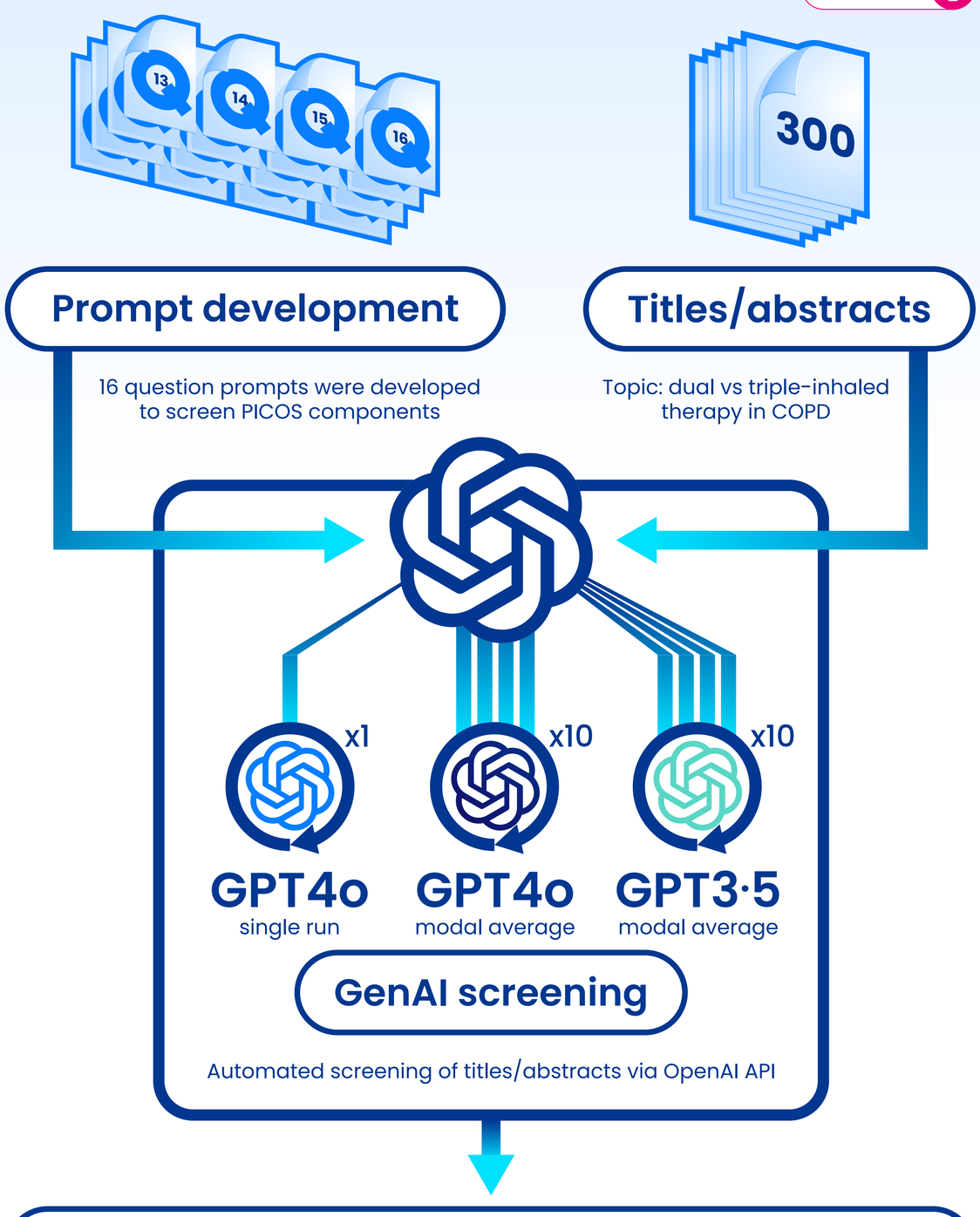
## Background

- Literature screening for systematic literature reviews (SLRs) is time-consuming
- GenAI may streamline this process, but there are limited data on its performance
- This study evaluated the level of agreement between genAI and a human reviewer when screening 300 titles and abstracts from a previous SLR (**Fig 1**)

## Method

**Fig 1.** Approaches to assess the performance of genAI vs human screening of literature

Study design




**Accuracy**
Overall agreement between AI and human decisions

**Recall**
The AI's ability to find all the relevant records that humans found

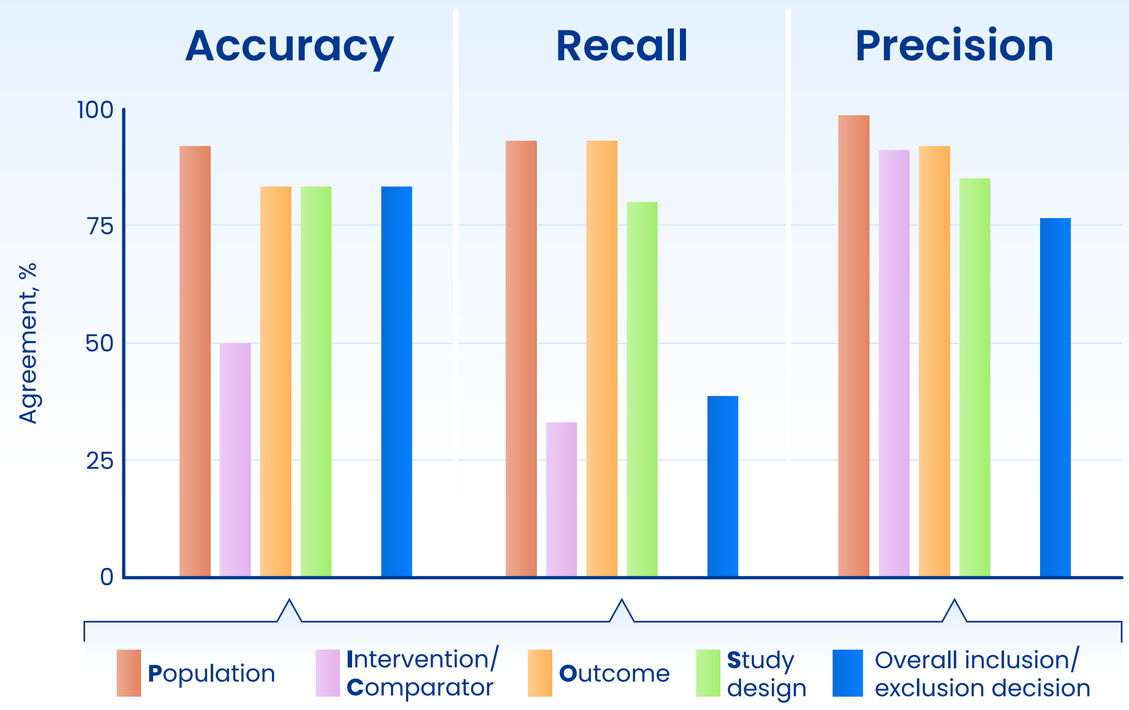
**Precision**
How many of the AI's identified records were relevant according to humans

## Results

- GenAI showed high agreement with the human reviewer on Population, Outcome and Study design, but low agreement on the Intervention/Comparator domain (**Fig 2**)
- Accuracy and precision were relatively high for the overall inclusion/exclusion decision, but recall was low

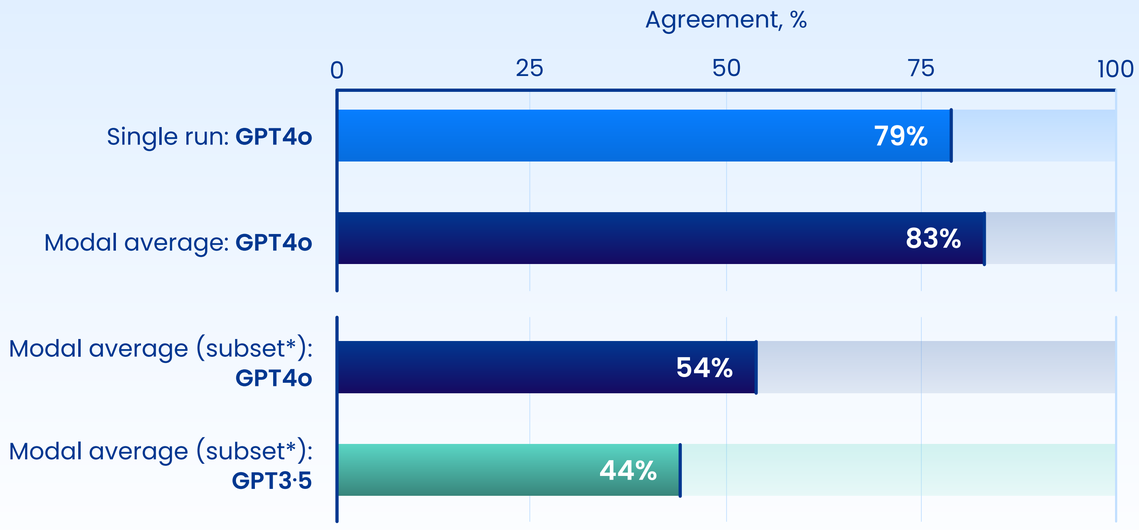
**Fig 2.** Agreement between single-run GPT4o and human reviewer across PICOS domains

Agreement




- Compared with the single-run approach, taking the modal answer of 10 runs with GPT4o resulted in slightly higher accuracy for the overall inclusion/exclusion decision (**Fig 3**)
- Conversely, accuracy was lower for the GPT3·5 subset than with the GPT4o subset

**Fig 3.** Overview of accuracy for overall inclusion/exclusion decision across model scenarios




*In a selected subset of 50/300 records

Model scenarios

- Overall, agreement between genAI and the human reviewer (accuracy) was similar to what would be expected between two humans[1]

**Authors:** George Miles,[1] Lucia Giles,[2] Bernard Kerr,[3] Brian Norman,[3] Gillian Sibbring[1]
**Author affiliation:** [1]Prime, Knutsford, UK; [2]Prime, Oxford, UK; [3]Prime, London, UK

**Abbreviations:** **AI**, artificial intelligence; **API**, application programming interface; **COPD**, chronic obstructive pulmonary disease; **GenAI**, generative AI; **PICOS**, population, intervention, comparator, outcome, study design; **SLRs**, systematic literature review; **GPT**, generative pre-trained transformer.

**Acknowledgments:** The authors thank Lloyd Gwishiri, MD, for writing assistance; Adam Moriarty and Joey Cusano, for creative input; Richard McDonald for editorial support, as well as the Prime Digital and Production teams for their support with development of this poster. All authors are employees of Prime, UK. All opinions are the authors' own and do not necessarily reflect those of our employer.

**Reference:** 1. Hanegraaf P, et al. *BMJ Open.* 2024;14(3):e076912.

**Presented at ISPOR Europe, Barcelona, Spain, Wednesday, 20 November 2024**